GUIDED HARMONIC SINUSOID ESTIMATION IN A MULTI-PITCH ENVIRONMENT

*Christine Smit and Daniel P.W. Ellis**

LabROSA, Electrical Engineering
Columbia University
New York NY 10025 USA
{csmit,dpwe}@ee.columbia.edu

ABSTRACT

We describe an algorithm to accurately estimate the fundamental frequency of harmonic sinusoids in a mixed voice recording environment using an aligned electronic score as a guide. Taking the pitch tracking results on individual voices prior to mixing as ground truth, we are able estimate the pitch of individual voices in a 4-part piece to within 50 cents of the correct pitch more than 90% of the time.

***Index Terms*—** pitch tracking, guided search, maximum a posteriori estimation

1. INTRODUCTION

We are interested in the dynamics of the voice and we would like to be able to use the large number of commercial recording available as evidence of vocal behavior. Unfortunately, most recordings have a mix of voices and instruments, which makes it difficult to focus on one voice. This work starts down the path of pulling out individual voices by looking at the problem of very exact pitch estimation. If we knew the exact pitch of the vocalist at any given time, we could largely filter out everything else.

The basic problem of single frequency estimation has been covered extensively. For example, instantaneous frequency estimations [1] and maximum likelihood estimation [2] have been used. However, musical notes generally consist of a harmonic pattern of frequencies. Linear predictive filtering has been used to tackle harmonic sinusoids[3], but this method still does not address the issue of multiple musical notes at once.

Finding multiple musical notes in a mix can be very difficult when we do not know what to expect. The truth is, however, that we often have pretty good information about which notes will be present at any time because we have some sort of electronic score, such as might be encoded in MIDI.

If we can align the score and the recording[4], we know almost exactly where to look for our notes. At this point, exact pitch estimation, which can follow vibratos and mistuned notes, should be fairly simple. We take a probabilistic approach to finding these exact pitch tracks using the score as a guide.

*This work was supported by the National Science Foundation (NSF) via grant IIS-0713334 and a fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

2. PHYSICAL SIGNAL MODEL

2.1. Time Domain Equations

A harmonic signal can be written as

$$x[n] = \sum_{i \in H} h_i[n] \tag{1}$$

where H is the set of harmonic numbers (say $\{1, 2\}$ for the first and second harmonic) we are using and

$$h_i[n] = A_i \cos\left(p \cdot i \cdot n + \phi_i\right), -\frac{N}{2} + 1 \leq n < \frac{N}{2}, \tag{2}$$

where A_i is the strength of harmonic i, p is the fundamental frequency in radians per sample, ϕ_i is the phase offset of harmonic i, and N is the window size. In our analysis, we assume, $h_i[n]$ simply repeats outside of the range $n \in \{-\frac{N}{2} + 1, \cdots \frac{N}{2}\}$, so $h_i[m] = h_i[m + N]$. In particular, we calculate all our Fourier transforms over the range $n = 0 \cdots N - 1$ (section 2.2).

To reduce side lobe effects, we use a Hann window,

$$w[n] = \frac{1}{2} \cdot \left(1 + \cos\left(\frac{2\pi n}{N}\right)\right), \tag{3}$$

so

$$x_w[n] = \sum_{i \in H} \left(w[n] \cdot h_i[n]\right) = \sum_{i \in H} h_{i,w}[n]. \tag{4}$$

2.2. Fourier Domain Equations

To simplify the problem of having multiple harmonics, we work with our model in the frequency domain, where each harmonic can be examined separately. We look over a range of Fourier coefficients,

$$K_i \in \{k_{0,i} - d, \cdots, k_{0,i}, \cdots, k_{0,i} + d\} \tag{5}$$

centered around the harmonic's nominal frequency given a nominal fundamental p_0

$$k_{0,i} = round\left(\frac{p_0 \cdot N \cdot i}{2\pi}\right), \tag{6}$$

where i is the harmonic number and d is some reasonable range around $k_{0,i}$. In this range, we assume that $X_w[K_i]$ is dominated by the one harmonic in question i.e. essentially equal to $H_i[K_i]$.

In the Fourier domain, calculated from $n = 0$ to $n = N - 1$, our harmonic signal is

$$H_i[k] = \frac{A_i}{2} \left(e^{j\phi_i} F_{\frac{N}{2}-1}\left(\frac{2\pi k}{N} - p \cdot i\right) + e^{-j\phi_i} F_{\frac{N}{2}-1}\left(\frac{2\pi k}{N} + p \cdot i\right)\right), \tag{7}$$

where

$$F_M(\theta) = 2\cos\left(\theta \cdot \frac{M}{2}\right) sinc_{M+1}\left(\frac{\theta}{2}\right) - 1 \quad (8)$$

and $sinc$ is a periodic sinc function,

$$sinc_M(\theta) = \frac{\sin(\theta M)}{\sin(\theta)}. \quad (9)$$

The Hann window has a simple 3-point Fourier Transform, so the windowed harmonic signal is simply

$$H_{w,i}[k] = \frac{1}{2}H_i[k] + \frac{1}{4}H_i[k-1] + \frac{1}{4}H_i[k+1] \quad (10)$$

2.3. Noise model

Because the real world has noise, we model the noise around each harmonic,

$$Y[K_i] = H_{w,i}[K_i] + N_{w,i}[K_i] \quad (11)$$

where $Y[K_i]$ is the Fourier transform of our full input around harmonic i and $N_{w,i}[K_i]$ is the noise around harmonic i.

Let us start with broad-band white noise in the time domain,

$$n_i[n] \sim N(0, \sigma^2_{n,i}). \quad (12)$$

In the Fourier domain, it is easier to notate $N_i[k]$ as a vector, $\boldsymbol{N}_i$, and we have that

$$Pr\left(\boldsymbol{N}_i\right) = Pr\left(\Re\left(\boldsymbol{N}_i\right)\right) \cdot Pr\left(\Im\left(\boldsymbol{N}_i\right)\right) \quad (13)$$

because the real and imaginary parts of $\boldsymbol{N}_i$ are independent. Furthermore,

$$Pr\left(\Re\left(\boldsymbol{N}_i\right)\right) \sim N\left(\boldsymbol{0}, \sigma^2_{n,i} \cdot \boldsymbol{\Sigma}_R\right) \quad (14)$$

and

$$Pr\left(\Im\left(\boldsymbol{N}_i\right)\right) \sim N\left(\boldsymbol{0}, \sigma^2_{n,i} \cdot \boldsymbol{\Sigma}_I\right), \quad (15)$$

where $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\Sigma}_R$ are both $N \times N$ dual diagonal matrices offset by one row and one column. Using zero-indexing, where k is the row and l is the column,

$$\boldsymbol{\Sigma}_R(k,l) = \begin{cases} N & \text{if } k = l = 0 \text{ or } \frac{N}{2} \\ \frac{1}{2}N & \text{if } k = l, k \neq 0, k \neq \frac{N}{2} \\ \frac{1}{2}N & \text{if } k = N - l, k \neq 0, k \neq \frac{N}{2} \\ 0 & \text{elsewhere} \end{cases} \quad (16)$$

and

$$\boldsymbol{\Sigma}_I(k,l) = \begin{cases} \frac{1}{2}N & \text{if } k = l, k \neq 0, k \neq \frac{N}{2} \\ -\frac{1}{2}N & \text{if } k = N - l, k \neq 0, k \neq \frac{N}{2} \\ 0 & \text{elsewhere.} \end{cases} \quad (17)$$

For the windowed noise, $\boldsymbol{N}_{w,i}$, the real and imaginary parts are similarly independent and we have that

$$Pr\left(\Re\left(\boldsymbol{N}_{w,i}\right)\right) \sim N\left(\boldsymbol{0}, \sigma^2_{n,i} \cdot \boldsymbol{B}\boldsymbol{\Sigma}_R\boldsymbol{B}^{\mathrm{T}}\right) \quad (18)$$

and

$$Pr\left(\Im\left(\boldsymbol{N}_{w,i}\right)\right) \sim N\left(\boldsymbol{0}, \sigma^2_{n,i} \cdot \boldsymbol{B}\boldsymbol{\Sigma}_I\boldsymbol{B}^{\mathrm{T}}\right) \quad (19)$$

where $\boldsymbol{B}$ is an $N \times N$ matrix,

$$\boldsymbol{B}(k,l) = \begin{cases} \frac{1}{2} & \text{if } k = l \\ \frac{1}{4} & \text{if } k = l - 1 \text{ or } l + 1 \pmod N \\ 0 & \text{elsewhere.} \end{cases} \quad (20)$$

Since we are looking at $N_{w,i}$ only around harmonic i, to calculate $Pr(N_{w,i}[K_i])$, we simply use the submatrices $\boldsymbol{\Sigma}_R(K_i, K_i)$ and $\boldsymbol{\Sigma}_I(K_i, K_i)$ as the covariances of the real and imaginary parts of the noise. Thus, our noise model is a narrow-band piece of broad-band white Gaussian noise.

3. MAXIMUM A POSTERIORI ESTIMATION

We calculate a maximum a posteriori (MAP) estimate of the pitch, p. Because we have no reasonable way of knowing in advance the harmonic strengths, $A_{i\in H}$, and phases, $\phi_{i\in H}$, we estimate these in addition the noise parameters, $\sigma_{n,i\in H}$. So,

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, Pr\left(\boldsymbol{\theta}|\boldsymbol{Y}, p_0\right), \quad (21)$$

where $\boldsymbol{\theta}$ is a vector of our parameters, [p, $A_{i\in H}$, $\phi_{i\in H}$, $\sigma_{n,i\in H}$], $\boldsymbol{Y}$ is the Fourier transform of our windowed input, and p_0 is the nominal pitch in our current window.

3.1. Defining probabilities

Using Bayes, we can say that

$$Pr\left(\boldsymbol{\theta}|\boldsymbol{Y}, p_0\right) \propto Pr\left(\boldsymbol{Y}|\boldsymbol{\theta}, p_0\right) Pr\left(\boldsymbol{\theta}|p_0\right). \quad (22)$$

For the prior, $Pr\left(\boldsymbol{\theta}|p_0\right)$, we simply want to capture the idea that the pitch of the current window, p, should be close to the nominal pitch from our score, p_0, so we assign

$$Pr\left(\boldsymbol{\theta}|p_0\right) \propto Pr\left(p|p_0\right) \sim N(p_0, \sigma^2_p) \quad (23)$$

and σ^2_p is the variance we expect around p_0.

Returning to equation 22, and assuming that $\boldsymbol{Y}$ only depends on the nominal pitch p_0 via the actual pitch p, we define

$$Pr\left(\boldsymbol{Y}|\boldsymbol{\theta}, p_0\right) = \prod_{i\in H} Pr\left(Y[K_i] \mid A_i, \phi_i, \sigma_{n,i}, p\right). \quad (24)$$

i.e. we combine the information from different harmonics by assuming that the conditional probabilities associated with each harmonic, $Pr\left(Y[K_i] \mid A_i, \phi_i, \sigma_{n,i}, p\right)$, are independent. Since the signal spectrum $H_i[K_i]$ is completely specified by the givens in equation 24, we can calculate $N_{w,i}[K_i] = Y[K_i] - H_i[K_i]$, so

$$Pr\left(Y[K_i] \mid A_i, \phi_i, \sigma_{n,i}, p\right) = Pr\left(N_{w,i}[K_i] \mid \sigma_{n,i}\right), \quad (25)$$

3.2. Finding the maximum

Finding the maximum of eqn. 22 directly is difficult, so we opted to search for a solution using the Nelder-Mead Simplex Method[5] as implemented by Matlab's fminsearch(), but any suitable maximization algorithm could be used. The main difficulty lies in seeding the search algorithm sufficiently close to the global optimum to avoid local maxima.

Our first step was to calculate approximate values for p and $A_{i\in H}$. We first observed that

$$H_i[K_i] \approx \frac{A_i}{2} F_{\frac{N}{2}-1}\left(\frac{2\pi k}{N} - p \cdot i\right) e^{j\phi_i}, \quad (26)$$

so, via eqn. 10 we have an expression for $|X_w[K_i]| \approx |H_{w,i}[K_i]|$ which is only a function of p and $A_{i\in H}$. We also observed that near the peak values of $|Y_w[k]|$,

$$|Y_w[K_{i,peak}]| \approx |X_w[K_{i,peak}]|, \quad (27)$$

where $K_{i,peak}$ consisted of the two largest values next to each other in $|Y_w[K_i]|$ for each harmonic i. Our initial estimates are thus a best fit of p and $A_{i\in H}$ to the approximation in eqn. 27. Initial phases ϕ_i were simply the phase of the peak value of $Y_w[K_i]$,

$$\phi_i \approx \angle Y_w[\underset{k}{\operatorname{argmax}}\, |Y_w[k \in K_{i,peak}]|]. \quad (28)$$

For the estimate of $\sigma_{n,i}$, we noted that in eqns. 18 and 19, the noise covariance scales linearly with $\sigma^2_{n,i}$. Our estimates of p, $A_{i\in H}$, and $\phi_{i\in H}$ were not always good enough to calculate

$N_w[K_i]$ accurately from $Y_w[k_i]$, particularly around the peaks. So, we estimated each $\sigma^2_{n,i}$ from the non-peak values of $Y_w[K_i]$,

$$\sigma^2_{n,i} \approx \frac{\sum_{i \in K_i \setminus K_{i,peak}} |N_w[i]|^2}{\sum_{i \in K_i \setminus K_{i,peak}} \left(\boldsymbol{B}\boldsymbol{\Sigma}_R\boldsymbol{B}^{\mathrm{T}}\right) + \sum_{i \in K_i \setminus K_{i,peak}} \left(\boldsymbol{B}\boldsymbol{\Sigma}_I\boldsymbol{B}^{\mathrm{T}}\right)}. \quad (29)$$

Thus, for a given window of the original signal, we derive initial estimates of p and $A_{i \in H}$ using eqns. 26 - 27, $\phi_{i \in H}$ from eqn. 28, and $\sigma_{n,i \in H}$ using eqn. 29. These values are passed to the optimizer to maximize eqn. 22, as defined in eqns. 23 - 25.

4. EXPERIMENTS

We first tested our algorithm on simulated data with two harmonics. We varied the power of each harmonic in the signal, while keeping the overall signal power constant. As shown in Figure 1, our algorithm was able to adapt to the varying harmonic power and maintain a good estimate of the fundamental frequency, and is unperturbed as energy is shifted between harmonics.

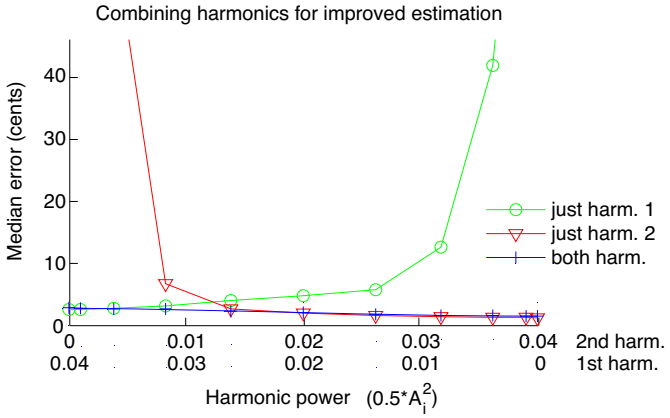

Figure 1: Fundamental tracking as power shifts between harmonics. At the left-most point, all signal power is in the fundamental, but by the right-most point, it has all transferred to the 2nd harmonic. Signal power, measured before windowing, is at $0.01\times$ (broadband) noise power. Plus symbols show result of estimation based on both harmonics, circle and triangle symbols show results based only on first and second harmonics, respectively.

We also tested our algorithm on a multi-track recording[6] of the opening to Guillaume de Machaut's Kyrie in Messe de Nostre Dame (c. 1365). Each of the four voices had been recorded individually and had been labeled with accurate start and stop times for each note. To obtain truth data for the pitch, we ran the well known fundamental frequency estimation algorithm YIN[7] over each individual track. For comparison purposes, we ran our own algorithm over the individual tracks as well as on the full mix. The YIN algorithm actually permits specification of a frequency search range, so we tried running YIN on the full mix, using the same guided search range that our algorithm used.

For these experiments, we chose a search range of 1 whole step below, and the same number of Hz above, the nominal pitch p_0. Our window length was 4096 samples (approximately 93 ms at 44.1 kHz) and our hop size was 1024 points for a 75% overlap between successive windows. This lead to about 1000 frames in notes per track. Our algorithm used the first and third harmonics for its frequency estimates.

As you can see in Table 1, our algorithm produced substantially the same results as YIN when they were both run on individual tracks, with the algorithms agreeing to within 50 cents (a quarter step) in 98.9% of windows. When our algorithm was run on the mix of voices, it was still able to find the correct pitch, within 50 cents, in more than 90% frames. The table also shows how YIN faired on the full mix, but told to only consider frequencies within a whole step of the nominal frequency. It was close to the actual frequency less than 50% of the time. We were frankly surprised that YIN did this well, considering that it is not designed to deal with multiple pitches at once. Finally, the table includes the results from naively guessing the nominal frequency of the note in each window, which differed from the tracked pitch by an average of about a quarter step – significantly larger than the results of our algorithm. Figure 3 shows histograms of these errors.

	RMSE (Hz)	RMSE (cents)	% < 50 cents
Prob (single)	0.743	4.77	98.9
Prob (mix)	4.15	24.8	91.4
YIN (mix)	31.1	162	45.9
f0	8.45	49.7	52.4

Table 1: Pitch tracking results on Machaut data, compared to YIN results on individual tracks. "Prob(single)" is the results of applying our probabilistic model to the individual tracks. "Prob(mix)" is the model applied to the 4-voice mixdown. "YIN(mix)" is YIN applied to the mixdown, and "f0" is the result taking the nominal (notated) pitch as the result.

Figure 2 illustrates the results of our experiments on one particular voice, the triplum line. As you can see in the top plot, YIN does an excellent job of following the pitch in the single track case; note the singer's vibrato at around 8 seconds. However, as the bottom plot shows, YIN cannot accommodate the mix of voices, even when given the correct frequency search range. In contrast, our algorithm generally finds a harmonic, which is mostly the correct fundamental.

5. DISCUSSION AND CONCLUSIONS

There are several novel advantages to our solution. First, unlike most pitch estimation algorithms, our solution is geared towards multiple simultaneous pitches. Secondly, our algorithm is able to adaptively weight the information it receives from different harmonics, depending on its local signal-to-noise ratio. A weak harmonic in a lot of noise affects our estimate less strongly than a strong harmonic with little noise. This results from the explicit optimization of the noise level in the vicinity of each harmonic.

A further advantage to our model is that it not only gives you a estimate of the fundamental frequency, it also gives you a goodness of fit measure in the probability calculations. Defining the solution in terms of a distribution also leaves the possibility open of using other tools from probability theory.

One limitation of our current algorithm is that it myopically searches for a single pitch in any given window even though we actually know approximately where interfering harmonics may lie. We suspect that we could improve further on our algorithm by using this knowledge either during the initial approximation stage or during the optimization stage.

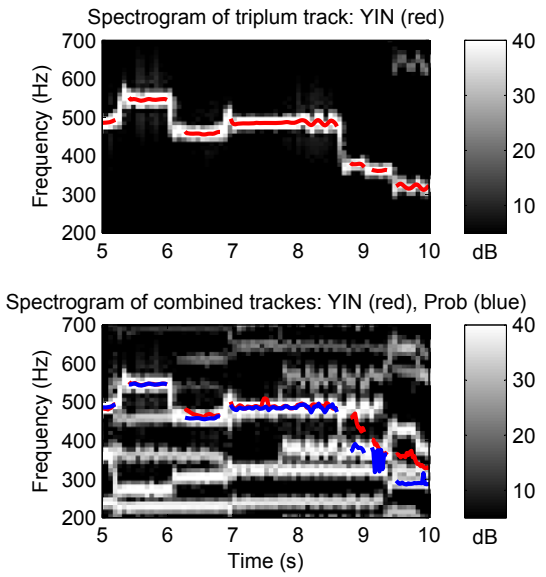

Figure 2: Spectrograms of 5 s excerpts from the multi-track recordings. The top plot shows a single track and the ground from YIN (red). The bottom plot shows a spectrogram of the mix with the full-mix YIN estimates (red) and our full-mix estimates (blue). Note the large difference between algorithms at around 9 s. Here, the correct fundamental is at about 365 Hz, but there is an interfering harmonic at about 320 Hz, which our algorithm jumps to.

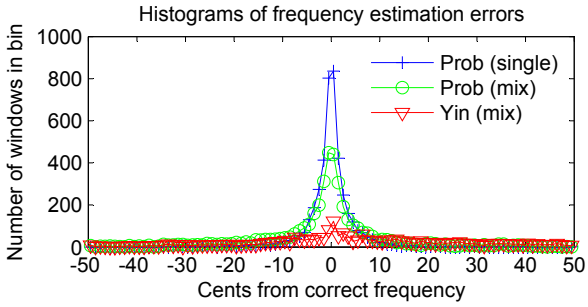

Figure 3: Histograms of the errors summarized in Table 1. Plus signs refer to differences between YIN and the probabilistic algorithm on individual voices. Circles show the errors when each line is tracked within the mix by the probabilistic algorithm. Triangles are the same data tracked by YIN. Substantial counts lie outside the ± 50 cent range shown in the plot.

Our algorithm is also quite slow. Because we do not have a direct solution to eqn. 21, we are forced to rely on an iterative algorithm to find the best solution. The speed of our algorithm is essentially a function of the speed of our optimizer and the goodness of our initial starting point.

Furthermore, the complexity of our algorithm grows with the number of harmonics we include in the estimate. Fortunately, the number variables over which we optimize eqn. 21 grows linearly with the number of harmonics. Unfortunately, even a linear increase in variables can cause the optimization routine to slow down considerably, which is why we have reported results with no more than two harmonics.

We believe that there are many other applications to this work. Obviously, there is no reason our model should not work for most other instruments. In fact, it should work wherever there is some way to calculate the constituent frequencies of a signal from the base frequency. So the slightly de-tuned harmonics found on the piano could easily be modeled. Furthermore, the accuracy of the pitch estimation lends itself to the study of many pitch-based phenomena such as tuning. We hope that the ability of this algorithm to recover detailed tuning nuances from recordings of real performances will make possible a broad new range of data-driven musicological investigation[1].

6. REFERENCES

[1] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[2] R. Kenefic and A. Nuttall, "Maximum likelihood estimation of the parameters of a tone using real discrete data," *Oceanic Engineering, IEEE Journal of*, vol. 12, no. 1, pp. 279–280, 1987.

[3] K. Chan and H. So, "Accurate frequency estimation for real harmonic sinusoids," *Signal Processing Letters, IEEE*, vol. 11, no. 7, pp. 609–612, 2004.

[4] R. Turetsky and D. Ellis, "Ground-truth transcriptions of real music from force-aligned midi syntheses," in *Proc. International Conference on Music Information Retrieval*, Baltimore, Oct. 2003.

[5] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder–mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998. [Online]. Available: http://link.aip.org/link/?SJE/9/112/1

[6] J. Devaney and D. P. Ellis, "An empirical approach to studying intonation tendencies in polyphonic vocal performances," *Journal of Interdisciplinary Music Studies*, vol. 2, no. 1-2, pp. 141–156, 2008.

[7] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002. [Online]. Available: http://link.aip.org/link/?JAS/111/1917/1

[1]The software can be downloaded from `http://www.ee.columbia.edu/~csmit/papers/waspaa_2009/`